DIU Génétique et Reproduction

Introduction à l'analyse NGS Exome

18 Novembre 2022

Philippe Dessen
Institut Gustave Roussy, Villejuif

dessen@igr.fr

http://pdessen.free.fr/KB

Abdelkader Heddar
Hôpital Bicêtre

abdelkader.heddar@aphp.fr

Bioinformatique

- Analyse de millions de « reads »
  - Lectures courtes de 32, 50, 100 voire 400 bases
  - Assemblage de novo (génomes inconnus)
  - Problèmes de « mapping » : localisation sur les génomes (par rapport à une référence connue). Cas du génome humain (versions évolutives depuis 2001 (hg19 en 2009) : 3.3 Gbases.(version révisée hg38 en 2013)

Bioinformatique

- Besoin d'algorithmes rapides mais aussi sensibles
  - Bowtie, BWA, BFAST, GEM, STAR ….
- Innovation dans l'indexation des génomes de référence (plus complexe que des index sur des mots simples (ref : blast suite)

Bioinformatique

Objectifs :

- Analyse des variations
  - Ponctuelles (polymorphismes ou mutations)
  - SNP de prédisposition
  - Mutations sur des gènes « »drivers » du cancer
  - Délétions, Gains (copy number)
  - Remaniements (duplications, translocations..)
- Expression des gènes
- Fusions de gènes

Cas des Exomes :
Capture des parties codantes

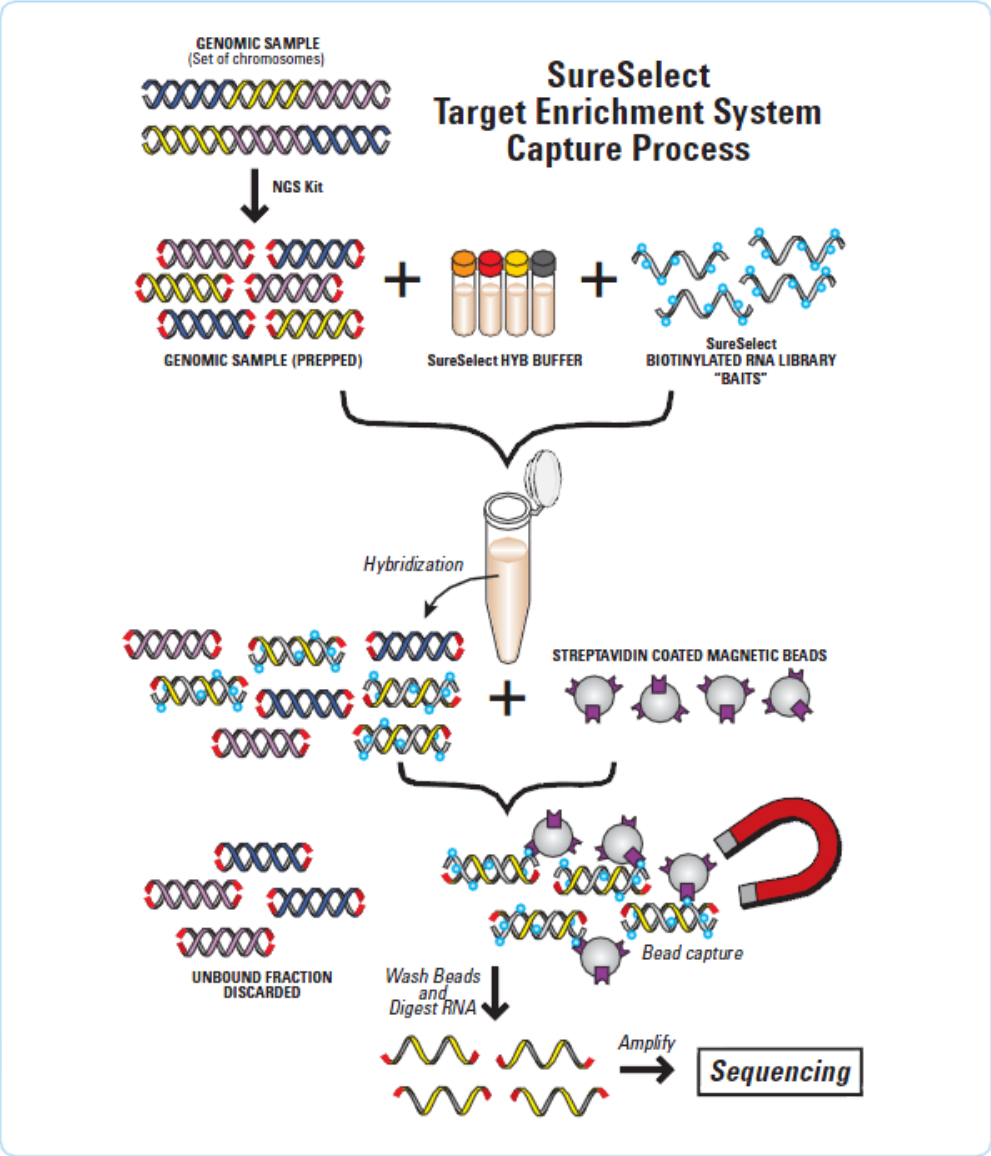


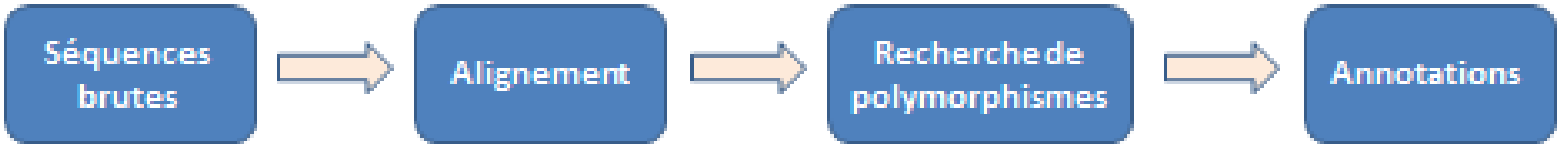
Figure 1. SureSelect Target Enrichment System Workflow

Permet de réduire les parties d'ADN à séquencer

Ex: exome :
Ensemble des zones du génome couvert par des séquences codant pour des gènes
Environ 50 Mbases (au lieu de 3.3 Gbases)

Capture sur des gènes ciblés (panel)

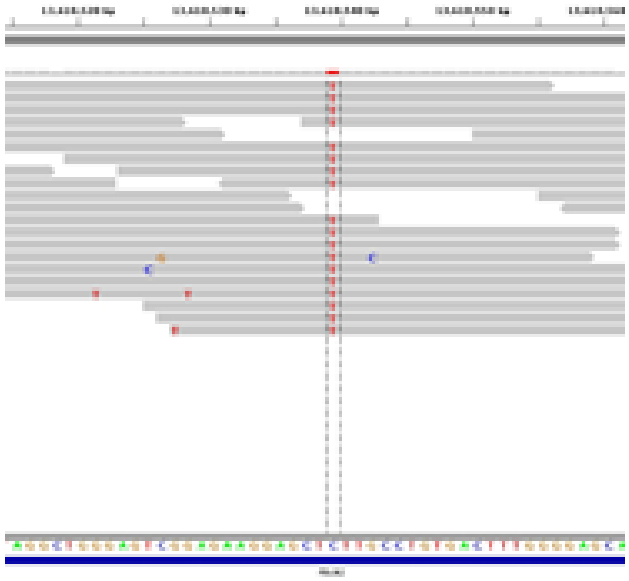
WorkFlow d'analyse de séquençage haut débit (recherche de polymorphismes)



- Outils d'alignement :
BWA

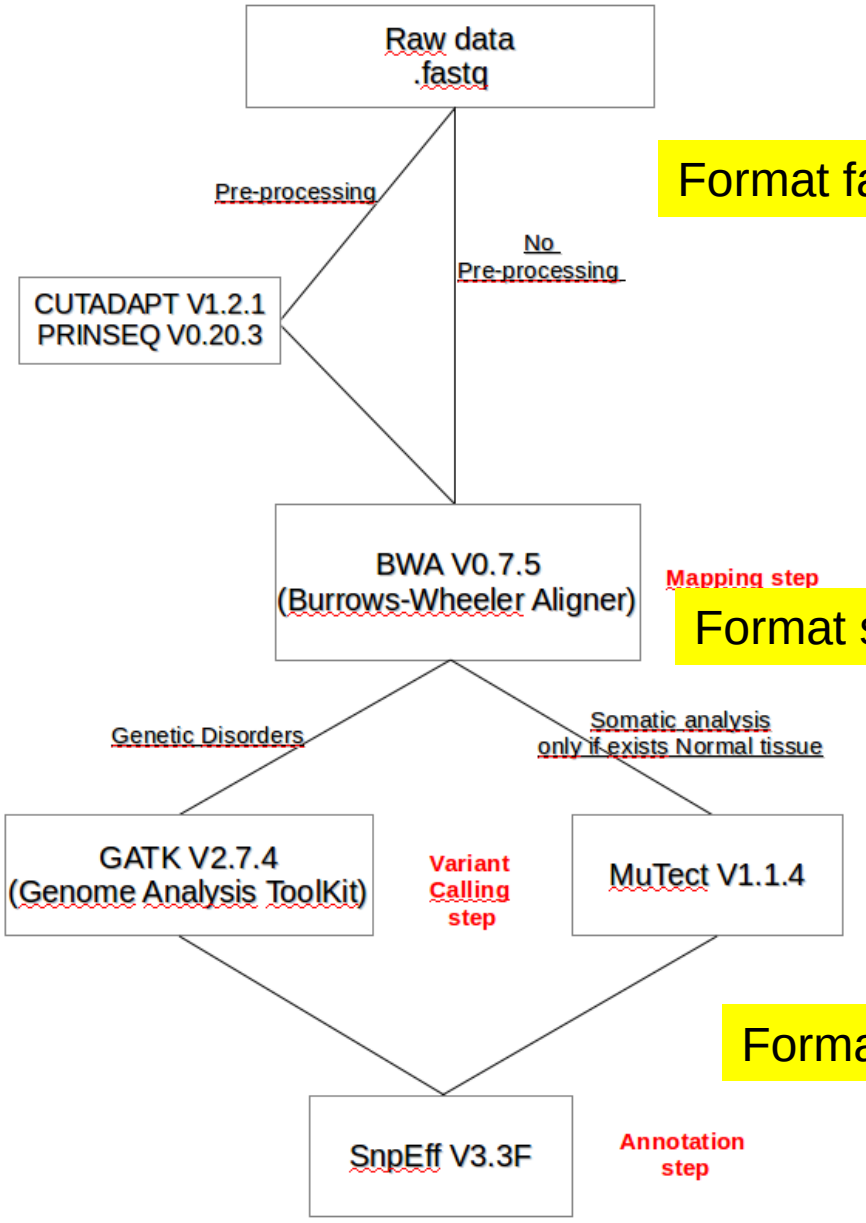
- Outils de recherche de polymorphismes :
Samtools – Varscan

- Outils d'annotations :
Scripts et programmes « in house » - Bases de données publiques – Sift



Visualisation IGV

Pipe Line DEVA (détection de variants)



Format fastQ

Preprocessing

suppression des adaptateurs
trimming

Mapping

Format sam/bam

Visualisation IGV

Variants caller

Format vcf

Annotation

Format FASTQ

Format [edit]

A FASTQ file normally uses four lines per sequence.

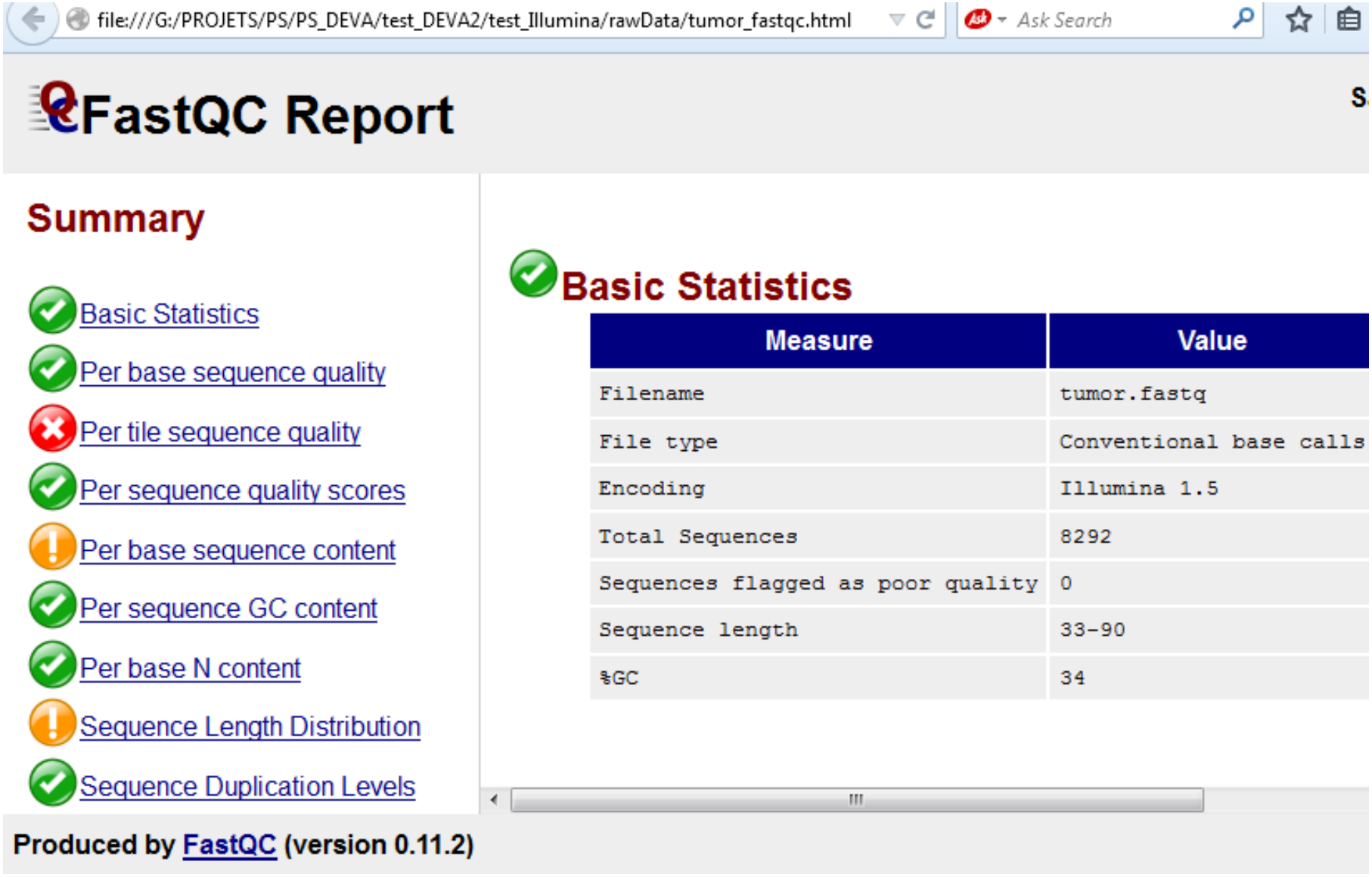
- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

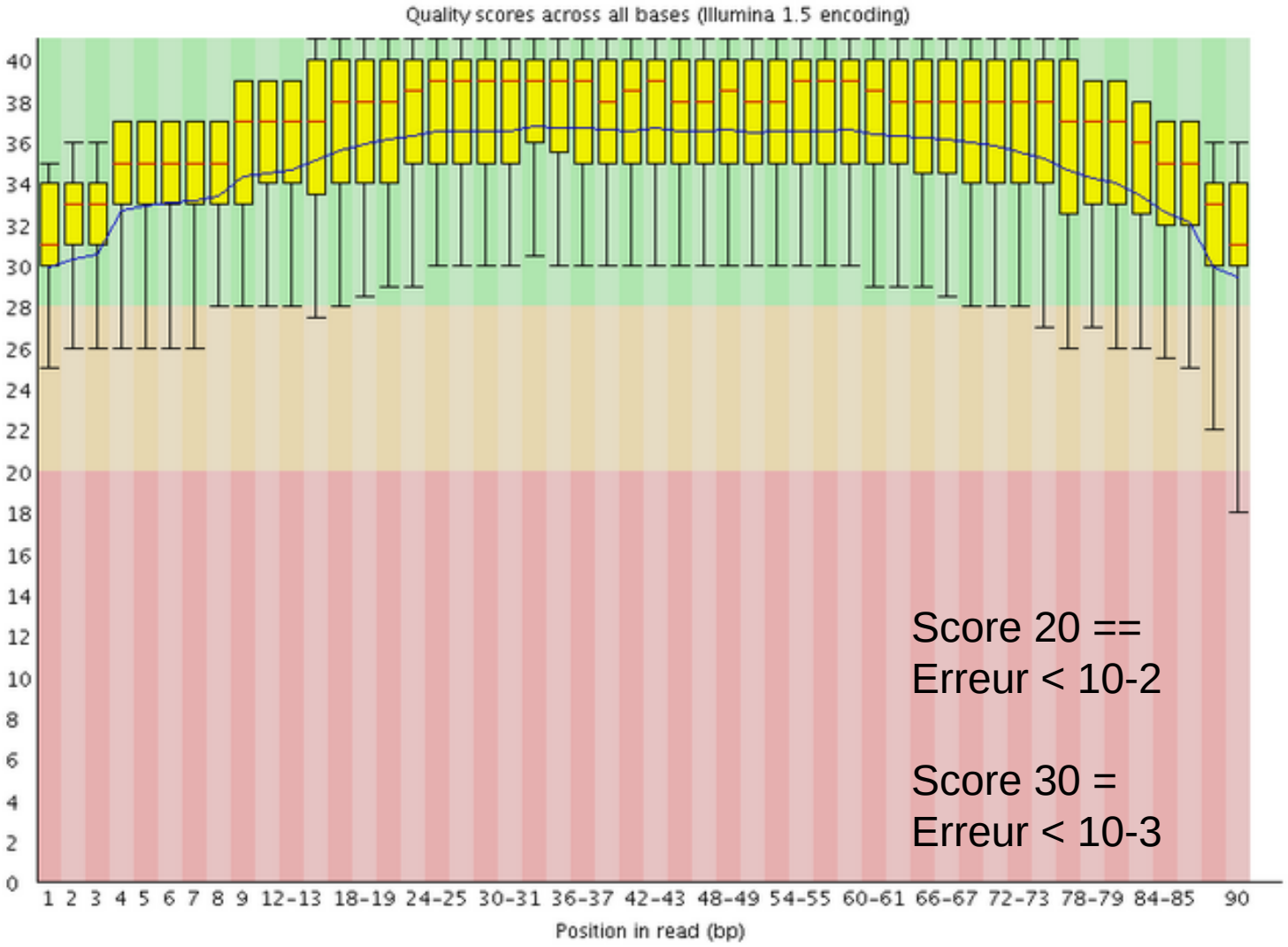
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

http://en.wikipedia.org/wiki/FASTQ_format


file:///G:/PROJETS/PS/PS_DEVA/test_DEVA2/test_Illumina/rawData/tumor_fastqc.html
Ask Search
FastQC Report
Summary
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Basic Statistics
Measure
Value
Filename
tumor.fastq
File type
Conventional base calls
Encoding
Illumina 1.5
Total Sequences
8292
Sequences flagged as poor quality
0
Sequence length
33-90
%GC
34
Produced by FastQC (version 0.11.2)

Per base sequence quality



Mapping avec BWA

Transformation d un fichier fastq en fichier sam (1 ligne par « read »)

1	@SQ	SN:ch	LN:249250621										
2	...												
3	@PG	ID:bw	PN:bw	VN:0.7.10-r7	CL:/share/apps//bwa-0.7.10/bwa mem /mnt/data/data/hg19/bwa_0.7.5/hg19.fa ../rawData/tu								
4	FCD1CCKACXX:5:1105:6376:82169	0	chr3	178866302	60	90M	*	0	0	CCGCTCTCCTCTC	bbbeeeeegg	NM:i:1	MD:Z:50C39	AS:i:85
5	FCD1CCKACXX:5:1105:6376:82169	0	chr3	178866383	60	90M	*	0	0	GCCCCGAGCGTG	a_a]aaacaccc	NM:i:0	MD:Z:90	AS:i:90
6	FCD1CCKACXX:5:2113:12893:41372	0	chr3	178867795	60	90M	*	0	0	TTTTTGGTGTTTT	__beeee`eeg	NM:i:0	MD:Z:90	AS:i:90
7	FCD1CCKACXX:5:2113:12893:41372	0	chr3	178867933	60	90M	*	0	0	ATCTGAAGAGGA	b]]bb]]]]YJG'	NM:i:0	MD:Z:90	AS:i:90
8	FCD1CCKACXX:5:1208:6968:60491	16	chr9	84266990	60	90M	*	0	0	CCCACCCTTATCA	___ceeecac^	NM:i:4	MD:Z:69G0C	AS:i:72
3186	FCD1CCKACXX:5:1306:11457:100061	0	chr3	178951975	60	90M	*	0	0	GCCAGAACTACA	Yb_]c]]cbZge	NM:i:0	MD:Z:90	AS:i:90
3187	FCD1CCKACXX:5:1211:6454:41169	0	chr3	178951979	60	90M	*	0	0	GAACTACAATCT	___eecccgac	NM:i:0	MD:Z:90	AS:i:90
3188	FCD1CCKACXX:5:1107:7413:8541	0	chr3	178951980	60	90M	*	0	0	AACTACAATCTT	___ceeeca`c'	NM:i:0	MD:Z:90	AS:i:90
3189	FCD1CCKACXX:5:1211:7662:30265	0	chr3	178951980	60	90M	*	0	0	AACTACAATCTT	BBBBBBBBBE	NM:i:0	MD:Z:90	AS:i:90
3190	FCD1CCKACXX:5:2104:2379:71551	0	chr3	178951983	60	90M	*	0	0	TACAATCTTTTGA	b`bZba^`ddd	NM:i:0	MD:Z:90	AS:i:90
3191	FCD1CCKACXX:5:1316:16055:58341	0	chr3	178951984	60	90M	*	0	0	ACAATCTTTTGA1	beeeeeggfd	NM:i:0	MD:Z:90	AS:i:90
3192	FCD1CCKACXX:5:2307:15835:57220	0	chr3	178951984	60	90M	*	0	0	ACAATCTTTTGA1	BBBBBBBgd^	NM:i:3	MD:Z:35C10	AS:i:78
3193	FCD1CCKACXX:5:1102:12228:68929	0	chr3	178951987	60	90M	*	0	0	ATCTTTTGATGAC	]T_c[ddbgfff	NM:i:0	MD:Z:90	AS:i:90
3194	FCD1CCKACXX:5:1106:18768:14288	0	chr3	178951987	60	90M	*	0	0	ATCTTTTGATGAC	bab_c_dgggg	NM:i:0	MD:Z:90	AS:i:90
3195	FCD1CCKACXX:5:2302:7290:8719	0	chr3	178951990	60	90M	*	0	0	TTTTGATGACATT	bbbeeeeegg	NM:i:1	MD:Z:54A35	AS:i:85
3196	FCD1CCKACXX:5:2112:17756:100625	0	chr3	178951991	60	90M	*	0	0	TTTGATGACATTC	^__`c^cdegga	NM:i:1	MD:Z:66T23	AS:i:85
3197	FCD1CCKACXX:5:2101:8014:68094	0	chr3	178951991	60	90M	*	0	0	TTTGATGACATTC	b]Yccd`^R]bb	NM:i:0	MD:Z:90	AS:i:90
3198	FCD1CCKACXX:5:1309:1737:6173	0	chr3	178951993	60	90M	*	0	0	TGATGACATTGC	^aZR_^\HHH	NM:i:0	MD:Z:90	AS:i:90
3199	FCD1CCKACXX:5:1104:4923:33417	0	chr3	178951994	60	90M	*	0	0	GATGACATTGCA	[JYacecee]ae	NM:i:1	MD:Z:41A48	AS:i:85
3200	FCD1CCKACXX:5:1212:2748:31984	0	chr3	178951995	60	90M	*	0	0	ATGACATTGCAT	___eeeceggc	NM:i:0	MD:Z:90	AS:i:90

tumor

Fichier sam (tabulé) en fichier bam (compressé)
Visualisable avec IGV (bam ET bam.bai)

Format SAM

1.4 The alignment section: mandatory fields

Each alignment line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

1 ligne par « read » avec position sur le genome (POS), qualité (MAPQ)
CIGAR : string qualifiant la comparaison avec le genome, Phred qualité » (QUAL).

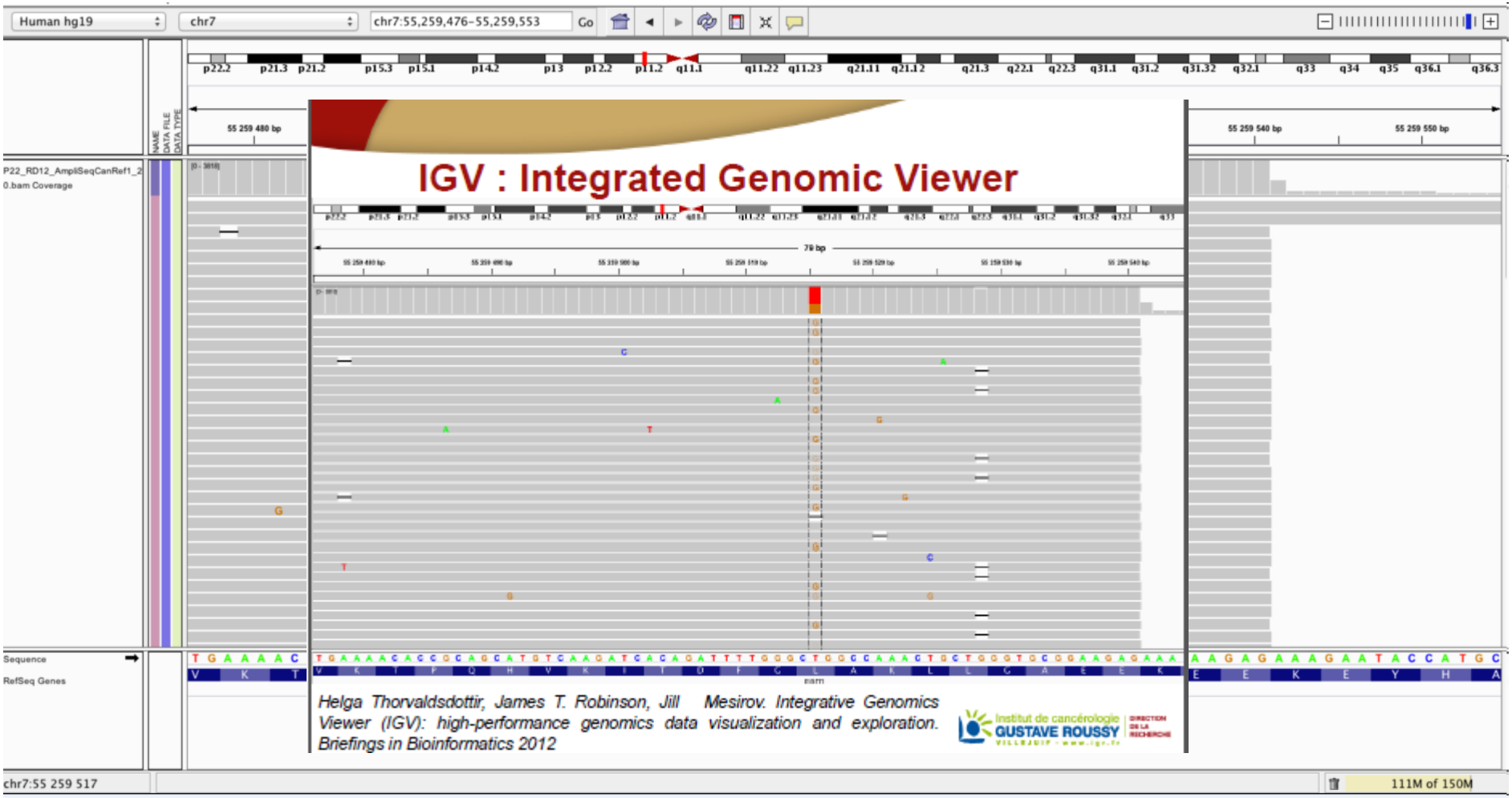
4. POS: 1-based leftmost mapping POSition of the first matching base. The first base in a reference sequence has coordinate 1. POS is set as 0 for an unmapped read without coordinate. If POS is 0, no assumptions can be made about RNAME and CIGAR.

5. MAPQ: MAPping Quality. It equals $-10\log_{10}\Pr\{\text{mapping position is wrong}\}$, rounded to the nearest integer. A value 255 indicates that the mapping quality is not available.

6. CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Détection de mutations sur une lignée cellulaire tumorale avec mutation EGFR L858R (30%)



AmpliSeq LifeTech (739 amplicons)
1 run IonTorrent (plaque 314 : 10 Mbases) : 376786 reads (88% match unique)
Profondeur de séquençage de 3795x
en chr7:55259516 T>G

Fichier mpileup control-tumor (ou control-disease)

1 ligne par position sur le génome

control tumorl

	A	B	C	D	E	F	G	H	I
	chr1	34173305	N	10	AAAAAAAAAA	^ceceaTTad	6	AAAAAA	ZecaZc
	chr1	34173306	N	10	CCCCCCCCCC	`cccgaT\cf	6	CCCCCC	^bec^c
	chr1	34173307	N	10	AAAAAAAAAA	Wcccc]NRca	6	AAAAAA	Y^eaYc
	chr1	34173308	N	10	CCCCCCCCCC	aaeccaZZch	6	CCCCCC	ac[ca_
	chr1	34173309	N	10	AAAAAAAAAA	aae_`cR]cg	6	AAAAAA	`ccc`a
	chr1	34173310	N	10	CCCCCCCCCC$	aceaee`eci	6	CCCCCC	iYc_i_
	chr1	34173311	N	1	G$	a	0	*	*
7	chr3	178952076	N	48	AAAAAAAAAAAAAAAAAAAAAAAAA	eafgegaeii[eihil	65	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	b_]`ec]aYeceggbih
8	chr3	178952077	N	49	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	ebggegYdiicbhil	63	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	`beca^aed^gcdifihX
9	chr3	178952078	N	51	GGGGGGGGGGGGGGGGGGGGGGGGG	bZeeeg`giihcfhi	63	GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	bbbcMbaecegcgifil
0	chr3	178952079	N	52	AAAAAAAAAAAAAAAAAAAAAAAAA	b_ede`\egidffiil	63	A$AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	b__cZbbee`gabieh
1	chr3	178952080	N	52	T$TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	bceeeeYeggddh	62	TTTTTTTTTTCTTTTTTTTTTTTTTTTTTTTTT	Zb[_d_eeYe`Hibhh
2	chr3	178952081	N	52	GGGGGGGGGGGGGGGGGGGGGGGGG	beeee^eggddiil	60	GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	\bbZeeQecZgagcgf
3	chr3	178952082	N	52	CCCCCCCCCCCCCCCCCCCCCCCCCCCC	beedc^eggddiig	61	C$CCCCCCCCCCCTCCCCCCCCCCCCCCCCCCCC	Z``]bbSec^eVggeec
4	chr3	178952083	N	52	AAAAAAAAAAAAAAAAAAAAAAAAA	baedc`eeg]eiide	60	A$AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	__ZbbSc^\cegeb^e
5	chr3	178952084	N	52	CCCCCCCCCCCCCCCCCCCCCCCCCCCC	bbddeceeec^ih	59	CCC$CCCCCCCCCCCCCCCCCCCCCCCCCCCC	]U__Pe[]g`gfcg`ghc
6	chr3	178952085	N	53	AAAAAAAAAAAAAAAAAAAAAAA	b_dd_^deec`ghi	57	GAAAAGGAGGAAAAAAAAAGAAAAAAAAAAAA	a^b_ageggecebf`\c
7	chr3	178952086	N	51	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	cd_`cecbJgiiigih	58	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	`^__de`ecaecdgc_d
8	chr3	178952087	N	49	C$CCCCCCCCCCCCCCCCCCCCCCCCC	_Pdeeb`ghhhbi	56	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	__cececWggefcZffg
9	chr3	178952088	N	49	AAAAAAAAAAAAAAAAAAAAAAAAA	^`ee`SgfhiXifihi	55	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	cecee^dcegc`gg`hY

cont_tum

Varscan2 ou VCL2HTML

Options de filtrage sur quality, %MAF non synonyme ….

1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

1.1 An example

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

1 ligne par position différente (ALT ; variant) par rapport à la référence (REF)

http://samtools.github.io/hts-specs/VCFv4.1.pdf

fichier VCF transformé en excel

1	chr	beg	rc_snp	REF	ALT	QUAL	AC	AF	AN	FS	DB	HaplScore	Dels	DP	MLEAC	MLEAF	MQ	MQ0	QD	GT	AB	AD	DP	GQ	MQ0	PL
2	chr1	16451413	rs1803527	C	T	127.03	2	1	2	0	DB	0	0	4	2	1	60	0	31.76	1-Jan		0,4	4	12	0	155,12,0
3	chr1	16451767	rs3754334	G	A	1056.77	1	0.5	2	7.415	DB	1.8662	0	118	1	0.5	60	0	8.96	0/1	0.48	57,61	63	99	0	1085,0,874
4	chr1	16456176	rs3768294	G	A	701.77	2	1	2	0	DB	0	0	19	2	1	60	0	36.94	1-Jan		0,19	19	57	0	730,57,0
5	chr1	16458722	rs116506614	C	T	1079.77	1	0.5	2	0	DB	4.8323	0	134	1	0.5	58.88	0	8.06	0/1	0.45	60,74	68	99	0	1108,0,989
6	chr1	16458814	rs2291805	C	T	1634.77	2	1	2	0	DB	0.8321	0	60	2	1	60	0	27.25	1-Jan		1,59	45	99	0	1663,129,0
7	chr1	16459507	rs11578289	T	A	75.77	1	0.5	2	0	DB	3.428	0	8	1	0.5	45.9	0	9.47	0/1	0.5	4,4	8	99	0	104,0,110
8	chr1	16459745	rs10907223	G	A	2964.77	2	1	2	0	DB	3.388	0	178	2	1	59.99	0	16.66	1-Jan		0,178	81	99	0	2993,226,0
9	chr1	16459832	rs55655135	C	T	1441.77	1	0.5	2	11.959	DB	3.539	0	267	1	0.5	60	0	5.4	0/1	0.48	129,138	86	99	0	1470,0,128
10	chr1	16460339	rs6669624	G	A	967.77	2	1	2	0	DB	0	0	32	2	1	59.71	0	30.24	1-Jan		0,32	29	78	0	996,78,0
11	chr1	16460541	rs6603855	T	C	624.77	2	1	2	0	DB	0.8667	0	21	2	1	60	0	29.75	1-Jan		0,20	19	48	0	653,48,0
12	chr1	16460743	rs33957854	TAAC	T	596.73	2	1	2	0	DB	9.5916		12	2	1	60	0	49.73	1-Jan		0,11	11	33	0	634,33,0
13	chr1	16460840	rs6603856	G	A	1844.77	2	1	2	0	DB	0.7887	0	59	2	1	60	0	31.27	1-Jan		1,58	50	99	0	1873,141,0
14	chr1	16461833	rs13375644	C	T	137.03	2	1	2	0	DB	0	0	4	2	1	60	0	34.26	1-Jan		0,4	4	12	0	165,12,0
15	chr1	16464260	rs11260742	T	C	852.77	2	1	2	0	DB	0	0	22	2	1	60	0	38.76	1-Jan		0,22	22	66	0	881,66,0
16	chr1	16464936	rs2291804	C	T	844.77	1	0.5	2	0	DB	0.9734	0	85	1	0.5	60	0	9.94	0/1	0.5	42,42	54	99	0	873,0,780
17	chr1	22139141	rs34297508	C	T	786.77	1	0.5	2	3.139	DB	2.2019	0	42	1	0.5	60	0	18.73	0/1	0.38	16,26	36	99	0	815,0,293
18	chr1	22141206	rs10917051	A	C	2202.77	2	1	2	5.572	DB	0.9789	0	108	2	1	59.9	0	20.4	1-Jan		2,106	62	99	0	2231,178,0
19	chr1	22147764	rs6684152	A	G	425.77	2	1	2	0	DB	0	0	11	2	1	60	0	38.71	1-Jan		0,11	11	33	0	454,33,0
20	chr1	22148164	rs6695528	G	A	1473.77	2	1	2	0	DB	0	0	56	2	1	59.98	0	26.32	1-Jan		1,55	39	99	0	1502,114,0
21	chr1	22148474	rs6687466	T	C	296.78	2	1	2	0	DB	0	0	9	2	1	60	0	32.98	1-Jan		0,8	9	24	0	325,24,0
22	chr1	22149935	rs897467	T	C	1699.77	2	1	2	0	DB	0.9967	0	96	2	1	59.91	0	17.71	1-Jan		0,96	50	99	0	1728,135,0
23	chr1	22150118	rs1138469	C	T	901.77	1	0.5	2	4.551	DB	0.7887	0	118	1	0.5	60	0	7.64	0/1	0.53	62,56	58	99	0	930,0,756
24	chr1	22150120	rs3736360	C	T	739.77	1	0.5	2	2.394	DB	0.7887	0	118	1	0.5	60	0	6.27	0/1	0.47	56,62	58	99	0	768,0,932
25	chr1	22150160	rs3736358	G	T	805.77	1	0.5	2	0	DB	5.2085	0	103	1	0.5	60	0	7.82	0/1	0.45	46,57	57	99	0	834,0,797
26	chr1	22154845	rs2228347	A	G	2698.77	2	1	2	0	DB	3.3892	0	157	2	1	60	0	17.19	1-Jan		1,156	73	99	0	2727,208,0

VCF : variant calling format

Les données dépendent des références en annotations

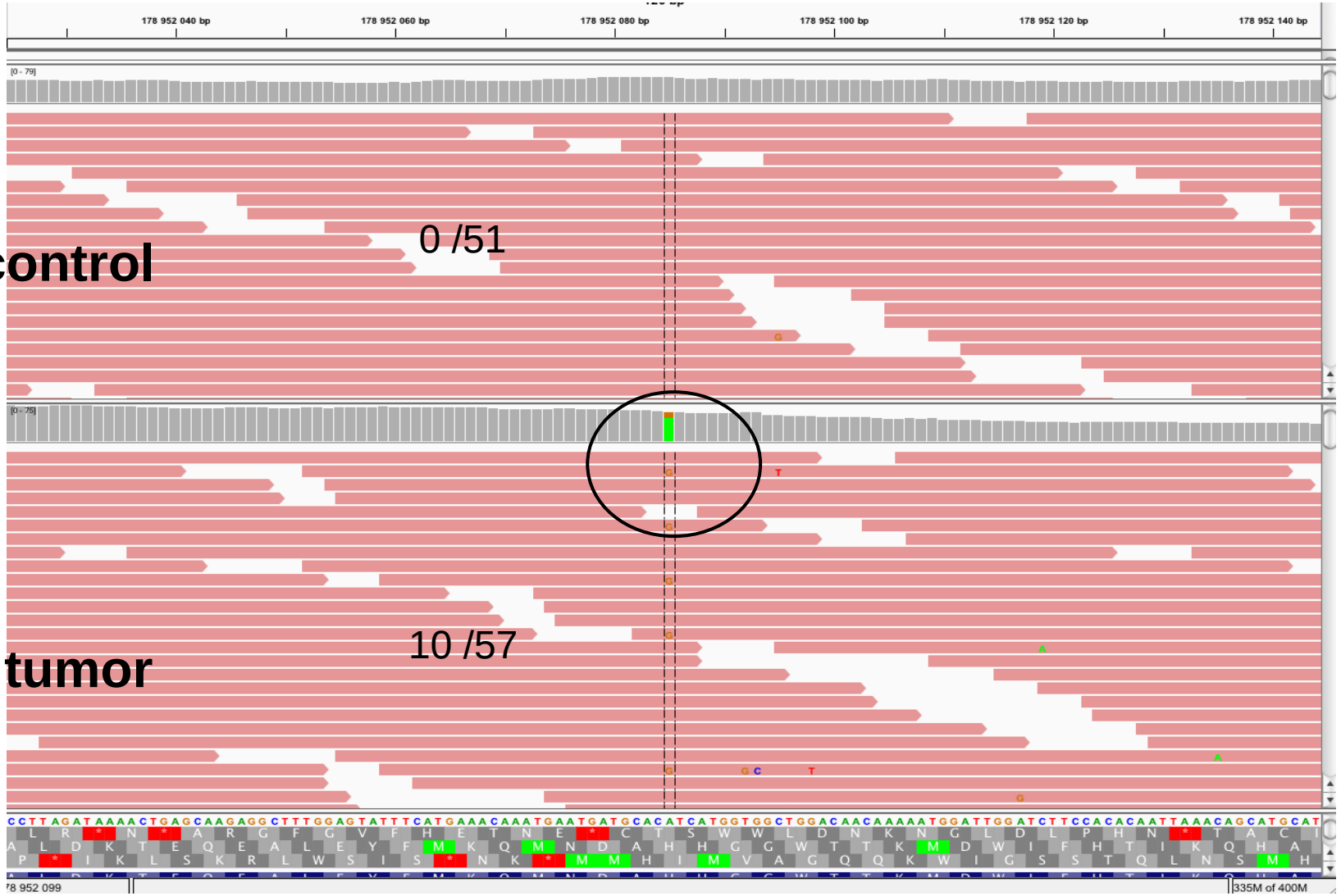
fichier VCF transformé en excel

chr	beg	EFFnb	EFFeffect	EFFimpact	EFFclass	EFFcodon	EFFaa	EFFlen	EFFsymb	EFFbiotype	EFFcoding	EFFtranscrit	EFFnex	EFFgenotype
chr1	16451413	7	UTR_3_PRIME	(MODIFIER				976	EPHA2	protein_coding	CODING	ENST00000358432	17	1
chr1	16451767	7	SYNONYMOUS_CODING	(LOW	SILENT	atC/atT	I958	976	EPHA2	protein_coding	CODING	ENST00000358432	17	1
chr1	16456176	12	INTRON	(MODIFIER				976	EPHA2	protein_coding	CODING	ENST00000358432	15	1
chr1	16458722	10	NON_SYNONYMOUS_CODING	(MODERATE	MISSENSE	cGg/cAg	R721Q	976	EPHA2	protein_coding	CODING	ENST00000358432	13	1
chr1	16458814	10	INTRON	(MODIFIER				976	EPHA2	protein_coding	CODING	ENST00000358432	12	1
chr1	16459507	12	INTRON	(MODIFIER				976	EPHA2	protein_coding	CODING	ENST00000358432	11	1
chr1	16459745	10	SYNONYMOUS_CODING	(LOW	SILENT	ctC/ctT	L661	976	EPHA2	protein_coding	CODING	ENST00000358432	11	1
chr1	16459832	10	SYNONYMOUS_CODING	(LOW	SILENT	ctG/ctA	L632	976	EPHA2	protein_coding	CODING	ENST00000358432	11	1
chr1	16460339	10	INTRON	(MODIFIER				976	EPHA2	protein_coding			9	1
chr1	16460541	10	INTRON	(MODIFIER				976	EPHA2	protein_coding			8	1
chr1	16460743	10	INTRON	(MODIFIER				976	EPHA2	protein_coding			8	1
chr1	16460840	10	INTRON	(MODIFIER				976	EPHA2	protein_coding			8	1
chr1	16461833	10	INTRON	(MODIFIER					EPHA2	processed_transcript			4	1
chr1	16464260	11	INTRON	(MODIFIER					EPHA2	processed_transcript			3	1
chr1	16464936	12	INTRON	(MODIFIER					EPHA2	processed_transcript			1	1
chr1	22139141	5	INTRON	(MODIFIER				272	LDLRAD2	protein_coding			1	1
chr1	22141206	19	NON_SYNONYMOUS_CODING	(MODERATE	MISSENSE	aAc/aCc	N134T	272	LDLRAD2	protein_coding			2	1
chr1	22147764	9	INTRON	(MODIFIER				272	LDLRAD2	protein_coding			3	1
chr1	22148164	9	INTRON	(MODIFIER				272	LDLRAD2	protein_coding			4	1
chr1	22148474	8	EXON	(MODIFIER					LDLRAD2	processed_transcript			2	1
chr1	22149935	9	SYNONYMOUS_CODING	(LOW	SILENT	tcA/tcG	S4350	4391	HSPG2	protein_coding			7	1
chr1	22150118	10	NON_SYNONYMOUS_CODING	(MODERATE	MISSENSE	Gtc/Atc	V4332I	4391	HSPG2	protein_coding			6	1
chr1	22150120	10	NON_SYNONYMOUS_CODING	(MODERATE	MISSENSE	aGc/aAc	S4331N	4391	HSPG2	protein_coding			6	1
chr1	22150160	10	SYNONYMOUS_CODING	(LOW	SILENT	Cgg/Agg	R4318	4391	HSPG2	protein_coding			6	1
chr1	22154845	10	SYNONYMOUS_CODING	(LOW	SILENT	gaT/gaC	D4104	4391	HSPG2	protein_coding			9	1

colonnes :
EFFeffect
EFFimpact
EFFclass
EFFcodon
EFFaa
EFFlen
EFFsymb
EFFbiotype
EFFcoding
EFFtranscrit
EFFnex
EFFgenotype

les colonnes dépendent de la ressource d'annotation

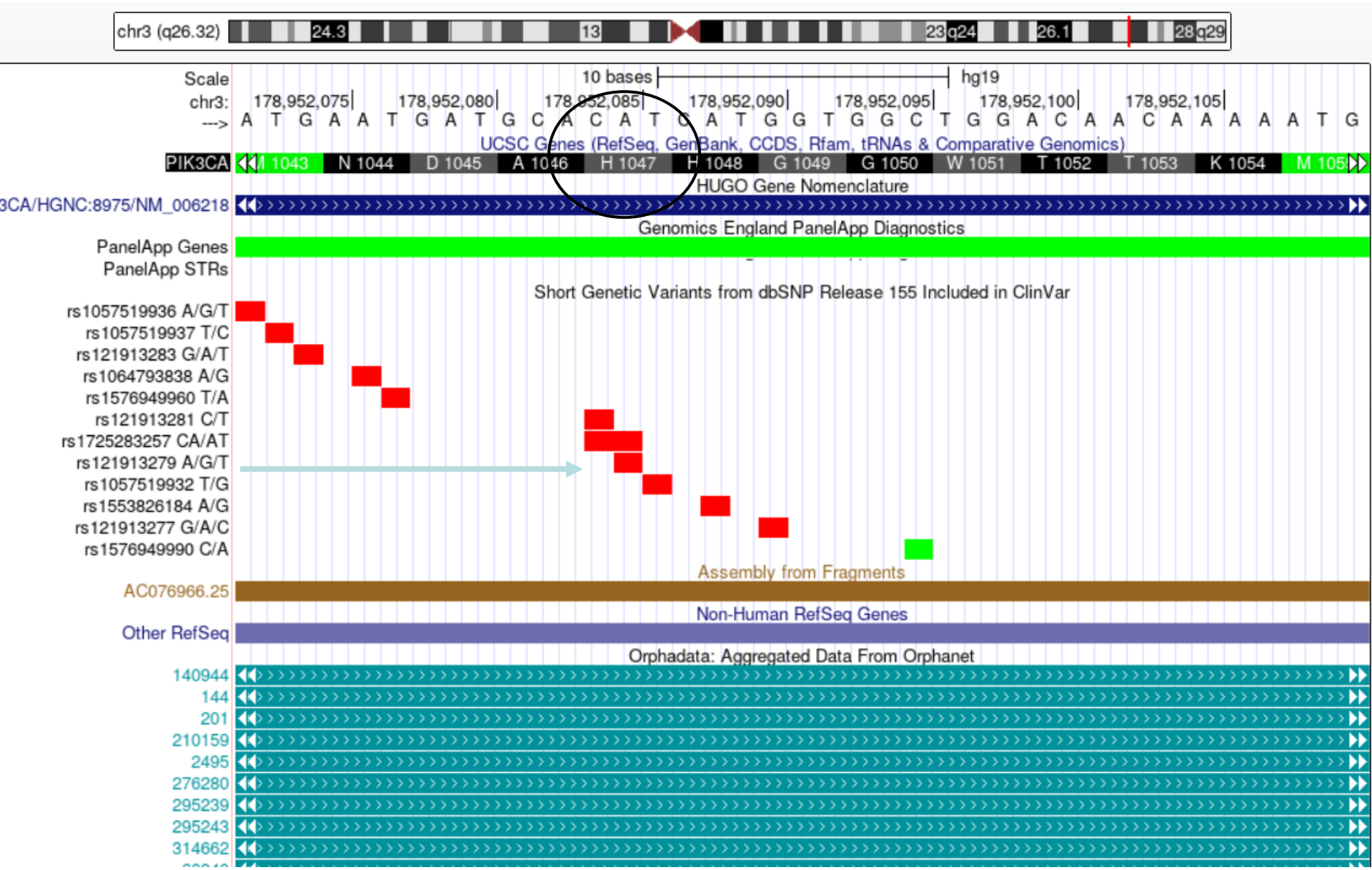
Mutation dans l'exon 21 de PIK3CA (cas d une tumeur)



chr3: 178952085 Mutation A > G (C**A**T > C**G**T) H 1047 R

3-178952085-A-G E p.His1047Arg missense Pathogenic

Localisation sur le GoldenPath (hg19)



ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

Disclaimer: These ACMG Standards and Guidelines were developed primarily as an educational resource for clinical laboratory geneticists to help them provide quality clinical laboratory services. Adherence to these standards and guidelines is voluntary and does not necessarily assure a successful medical outcome. These Standards and Guidelines should not be considered inclusive of all proper procedures and tests or exclusive of other procedures and tests that are reasonably directed to obtaining the same results. In determining the propriety of any specific procedure or test, the clinical laboratory geneticist should apply his or her own professional judgment to the specific circumstances presented by the individual patient or specimen. Clinical laboratory geneticists are encouraged to document in the patient's record the rationale for the use of a particular procedure or test, whether or not it is in conformance with these Standards and Guide-

Analyse de Whole Genome (WGS) ou d'exomes (WES)

Analyses de variants structuraux (copy number, translocations ….)

https://www.ncbi.nlm.nih.gov/dbvar/content/overview/
https://en.wikipedia.org/wiki/Structural_variation_in_the_human_genome

BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 15 2010, pages 1895–1896
doi:10.1093/bioinformatics/btq293

Genome analysis

SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data

Bruno Zeitouni[1,2,3,*], Valentina Boeva[1,2,3,4], Isabelle Janoueix-Lerosey[1,4], Sophie Loeillet[1,5], Patricia Legoix-né[1], Alain Nicolas[1,5], Olivier Delattre[1,4] and Emmanuel Barillot[1,2,3]

[1]Institut Curie, [2]INSERM, U900, Paris F-75248, [3]Mines ParisTech, Fontainebleau F-77300, [4]INSERM, U830 and [5]CNRS, UMR3244, Université Pierre et Marie Curie, Paris F-75248, France

Associate Editor: Alex Bateman

ABSTRACT

Summary: We present SVDetect, a program designed to identify genomic structural variations from paired-end and mate-pair next-generation sequencing data produced by the Illumina GA and ABI SOLiD platforms. Applying both sliding-window and clustering strategies, we use anomalously mapped read pairs provided by current short read aligners to localize genomic rearrangements and classify them according to their type, e.g. large insertions–deletions, inversions, duplications and balanced or unbalanced inter-chromosomal translocations. SVDetect outputs predicted structural variants in various file formats for appropriate graphical visualization.

Availability: Source code and sample data are available at http://svdetect.sourceforge.net/

Contact: svdetect@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 17, 2010; revised on May 4, 2010; accepted on June 1, 2010

GASV (Sindi *et al.*, 2009), BreakDancer (Chen *et al.*, 2009) and others (see review, Medvedev *et al.*, 2009).

Here, we present a new freely available program called SVDetect for SV detection and type prediction from PEM data. SVDetect identifies different types of SVs, e.g. large insertions–deletions and inversions, with both clustering and sliding-window strategies, and helps to visualize them at the genomic scale. Compared to other tools, the novelty of our method consists in its multiple ability to: (i) analyze both paired-end and mate-pair sequencing data; (ii) use unique PEM constraints to improve SV detection; (iii) predict various types of tandem duplication and to distinguish between balanced and unbalanced rearrangements; (iv) compare SVs across multiple samples; (v) construct copy number profiles; and (vi) create various output file formats for graphical views of SV.

2 METHODS

The first step in SVDetect is to regroup all pairs that are suspected to originate from the same SV. The input consists of paired-ends mapped to the

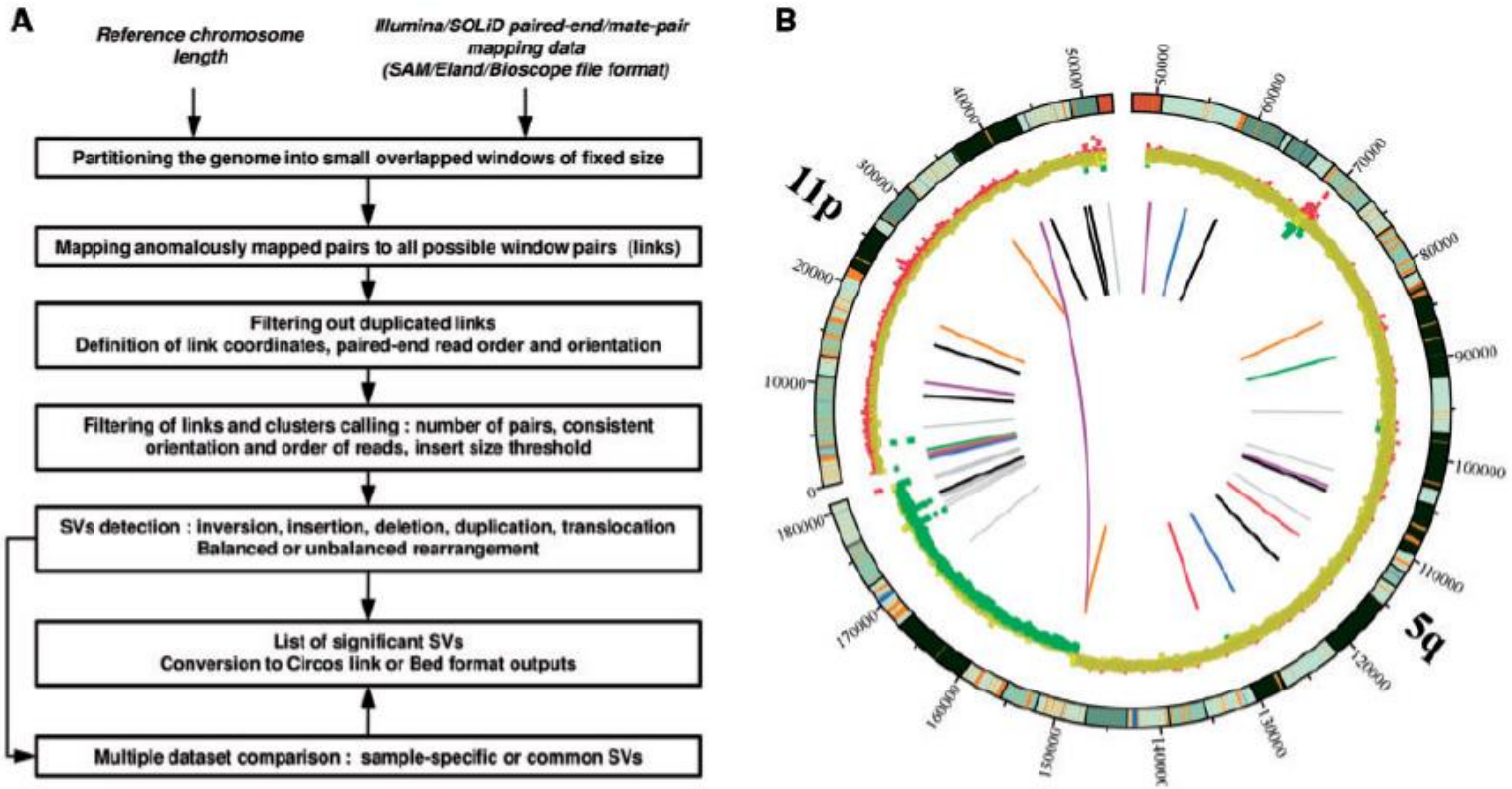
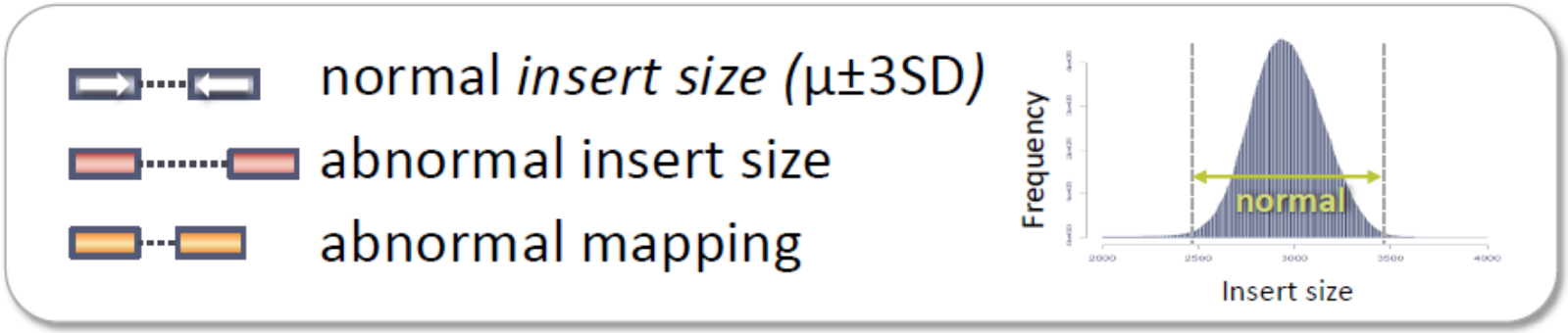
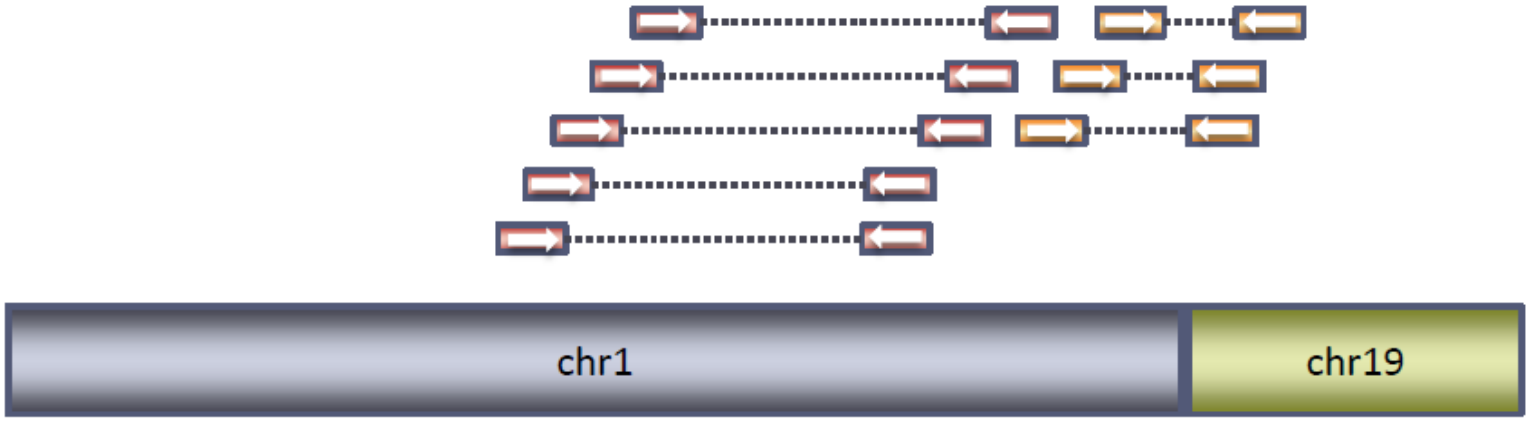


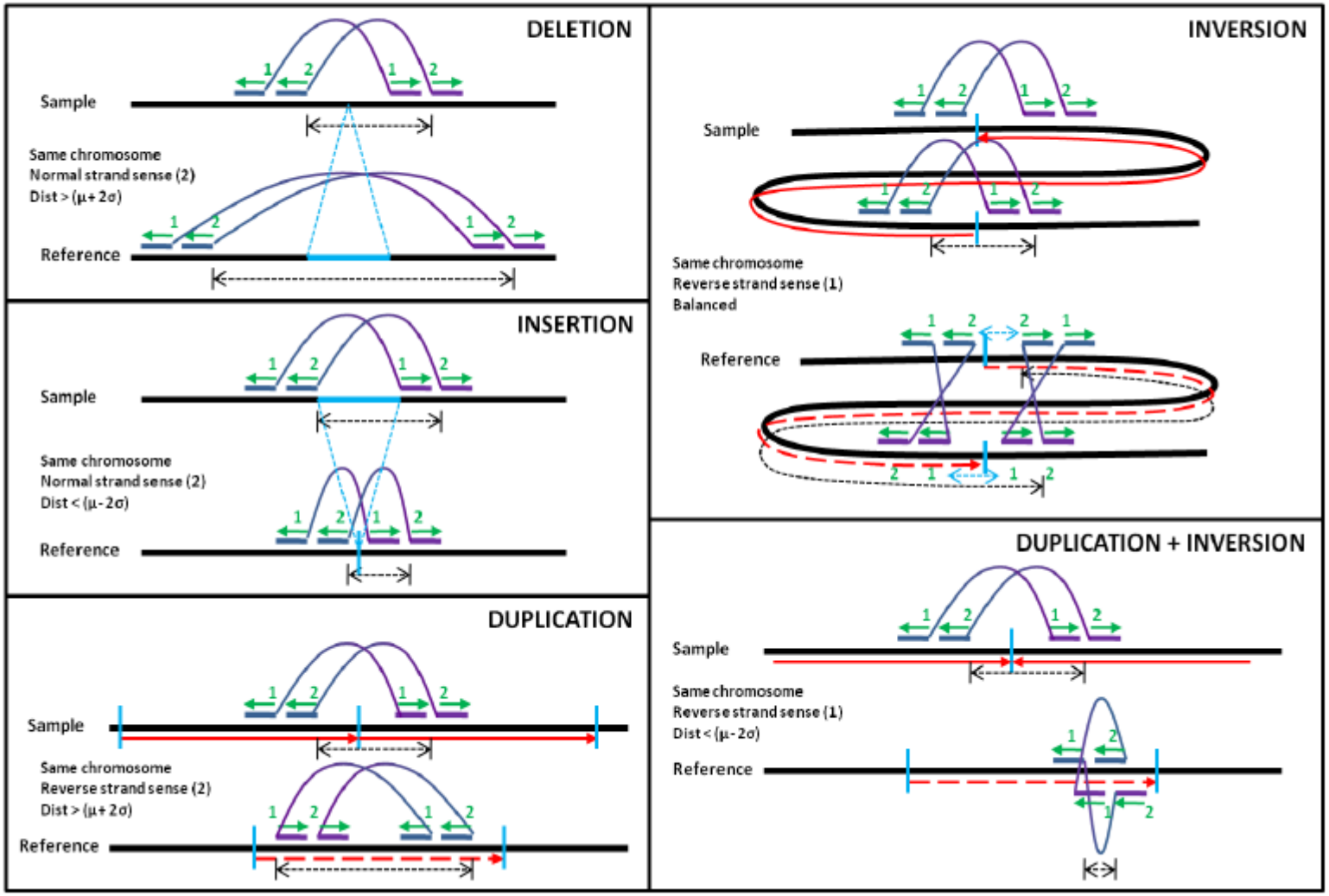
Fig. 1. Overview of SVDetect algorithm and output. (**A**) The workflow. (**B**) Graphical visualization of predicted SVs. Genomic locations of inter- and intra-chromosomal links are shown using the Circos software. Starting from outside of the circle, the following features are displayed: chromosome ideograms, scatter plot of the copy-number profile and color-coded spans of chromosomal links.

Séquençage par pair end (insert ~ 200-300 bases)
Détection des paires anormales





Intra-chromosomal SVs



Inter-chromosomal SVs